(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number: 0 167 364 A1

(12) EUROPEAN PATENT APPLICATION

(21) Application number: 85304627.4

(22) Date of filing: 28.06.85

(51) Int. Cl.⁴: G 10 L 3/00

(30) Priority: 06.07.84 US 628583

(43) Date of publication of application: 08.01.86 Bulletin 86/2

(84) Designated Contracting States: DE FR

(71) Applicant: AMERICAN TELEPHONE AND TELEGRAPH COMPANY
550 Madison Avenue
New York, NY 10022(US)

(72) Inventor: DonVito, Marc Bernard
4 Lancelot Court Apt. 12
Salem New Hampshire 03079(US)

(72) Inventor: Schoenherr, Brian William
1038 Fellsway
Medford Massachusetts 02155(US)

(74) Representative: Watts, Christopher Malcolm Kelway, Dr.
Western Electric Company Limited 5, Mornington Road
Woodford Green Essex, IG8 0TU(GB)

(54) Speech-silence detection with subband coding.

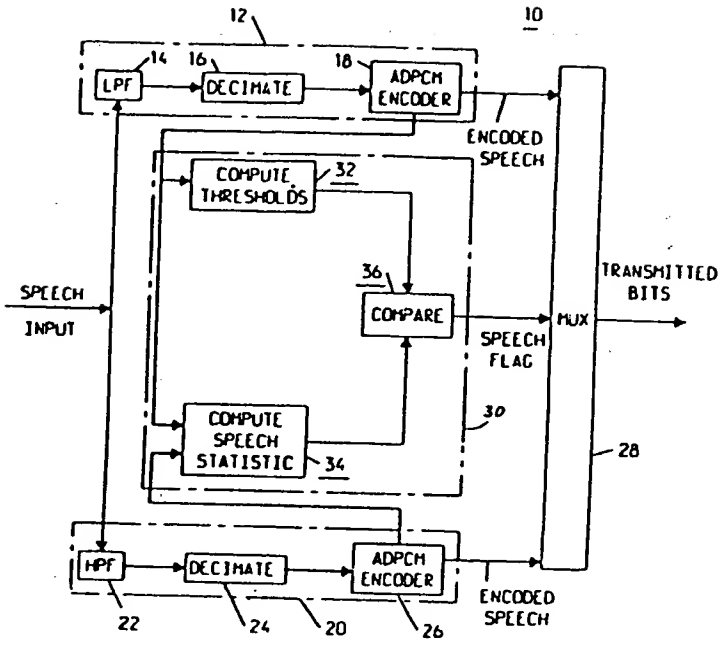
(57) Speech detection is accomplished in conjunction with two-band subband encoding. A detection statistic T (i_{τ_0}), used to estimate the short-term speech energy, is developed from energy estimates made in each subband. A speech presence energy threshold λ_{ON} a speech silence energy threshold λ_{OFF} and λ_{OFF} are computed which adapt to the long-term speech level. The detection statistic is compared to the thresholds to make a decision concerning the presence or absence of speech.

Also disclosed are considerations for extrapolating the detection to result in an arrangement with more than two subbands.

EP 0 167 364 A1

./...

FIG. 1


12
10
LPF
14
16
DECIMATE
18
ADPCM ENCODER
ENCODED SPEECH
COMPUTE THRESHOLDS
32
36
COMPARE
SPEECH INPUT
MUX
TRANSMITTED BITS
SPEECH FLAG
30
COMPUTE SPEECH STATISTIC
34
28
HPF
DECIMATE
ADPCM ENCODER
ENCODED SPEECH
22
24
20
26

SPEECH-SILENCE DETECTION WITH SUBBAND CODING

Technical Field

The invention relates to signal processing generally, and more particularly to means for detecting intervals of silence in encoded speech.

Background of the Invention

Normal human speech includes intervals of silence which will be referred to herein as "speech silence." When the speech is transmitted electronically, such as in a communications network, the speech-silence occupies a significant portion of the total transmission time. This leads to inefficient use of the communications network, since the only information which is transmitted during the course of the entire speech-silence interval, no matter how long, is the existence of the interval and its duration.

Efforts have been made to improve the efficiency of transmission by inserting other information, such as data, in the silence intervals on a time assignment basis. Such an approach is presently used for transatlantic cable and satellite communications which are known as TASI (time assignment and speech interpolation) systems. A system of this type is described, for instance in U.S. Pat. 4,100,377.

Speech silence may be detected even in voice signals which have already been digitally encoded into a pulse code modulated (PCM) format. This is described, for example, in U.S. Pats. 3,909,532 and 4,449,190.

Where both encoded speech and data signals share a carrier on a time assignment basis, there is a need for a high degree of accuracy in the determination of speech-silence intervals in order to permit the maximum use of the interval without degradation of the reconstructed speech. Of primary interest in this regard, therefore, are speech-silence boundaries. These are a transition either from voice to silence or from silence to voice. Accordingly,

there is a need for speech-silence boundary detection with improved accuracy.

Summary of the Invention

In accordance with the novel method and apparatus of the present invention, speech-silence boundaries are detected in the digitally encoded data of at least two subbands of the speech signal. Energy estimates are made for each of the frequency subbands for generating a detection statistic to estimate short-term speech energy. A threshold which is adapted to the long-term speech level is computed. This threshold is compared to the detection statistic to make a decision as to the presence of a silence interval. The resulting detection has significantly improved accuracy over detection using only one frequency band.

Brief Description of the Drawing

FIG. 1 is a functional block circuit diagram of a two-band subband encoder with speech detection in accordance with one example of the present invention.

FIG. 2 is a functional flow diagram showing in more detail a speech statistic computation subunit of the apparatus of FIG. 1.

FIG. 3 is a functional flow diagram showing in more detail a threshold computation subunit of the apparatus of FIG. 1.

FIG. 4 is a functional flow diagram showing in more detail a speech determination subunit of the apparatus of FIG. 1.

Detailed Description

The two-band subband encoder 10 with speech detection shown in FIG. 1 includes a lower frequency subband, or low band encoding circuit 12 made up of a low pass quadrature mirror filter 14, a by-two decimator 16, and an ADPCM (adaptive digital pulse code modulation) encoder 18. In parallel with the low band circuit 12 is a higher frequency subband, or high band encoding circuit 20 made up of a high pass quadrature mirror filter 22, a by-

two decimator 24, and an ADPCM encoder 26. Both of the encoding circuits 12, 20 operate with a sampling rate of 12 kHz (kilohertz) and receive the same 5.5 kHz analog speech input signal. They send their outputs to a multiplexer 28 for transmission. The details of subband encoding circuits such as the circuits 12, 20 and the multiplexer 28 are known to those in the art and are described, for example, in the U.S. Pat. 4,048,443 in "Sub-band Coding," by R. E. Crochiere in the Bell System Technical Journal, vol. 60, No. 7, Part 2, pp. 1633-1653, Sept. 1981, and in "Digital Voice Storage In a Microprocessor," by J. L. Flanagan, J. D. Johnston, and J. W. Upton, IEEE Transactions On Communications, Feb. 1982, vol. COM 30, no.2, pp.336-345.

A speech detector 30, which includes a speech threshold computing subunit 32, a speech statistic computing subunit 34, and a determining subunit 36 is adapted to provide an output to the multiplexer 28 which will result in the insertion of a speech presence indicator, or speech flag, in the transmitted output. The input to the speech threshold computing subunit 32 is the step size information from the low band encoder 12. The input to the speech statistic computing subunit 34 is the sample step size information from both the low band encoder 12 and the high band encoder 20. Both the threshold subunit 32 and the statistic subunit 34 give their output to the speech determining subunit 36.

The statistic computing subunit 34 is shown in greater detail in FIG. 2. Speech detection is accomplished by deriving information from the encoders 12, 20 and using it to determine whether speech is present or absent. Each of the encoders 12, 20 in the course of its normal encoding function makes a separate determination of the quantizer step size, based on the signal amplitude in its respective subband. For computational efficiency, the log of the step size is determined and used as a pointer to a step-size table. The log step-size parameters are used as estimates of the speech in each band at a given time.

Referring now to FIG. 2, the speech sampling period is represented by τ_0. The log of the step size in the low band is represented by $d_L(i\tau_0)$, while the log of the step size in the high band is represented by $d_H(i\tau_0)$ at time $t=i\tau_0$. Let $T(i\tau_0)$ be the speech detection statistic used to determine the speech level. Let σ_L and σ_H be fixed weights associated with $d_L(i\tau_0)$ and $d_H(i\tau_0)$, and let β_{DS} be a fixed weight such that $0<\beta_{DS}<1$. Then a detection statistic $T(i\tau_0)$ can be computed as follows:

$$T(i\tau_0) = \beta_{DS}T[(i-1)\tau_0] + \sigma_L d_L(i\tau_0) + \sigma_H d_H(i\tau_0). \quad (1)$$

The detection statistic $T(i\tau_0)$ is smoothed to become a low-pass filtered sum of speech information taken from each subband. The weight β_{DS} is chosen to give $T(i\tau_0)$ a specific time constant which controls the necessary smoothing of the information. A time constant of 16 milliseconds has been found to be suitable. The constants σ_L and σ_H determine the relative weight given to each subband. It has been found to be particularly advantageous to set σ_H at a value of about 1.5 to 2 times the value of σ_L. This accentuates discrimination in the high subband, which contains more information for the detection of fricatives and other consonants. The values of these constants for a particular application may be readily determined by means of laboratory tests by one skilled in the art.

FIG. 3 shows the method of computing a speech presence energy threshold λ_{ON} and a speech silence energy threshold λ_{OFF}. This method is very similar to that used in ADPCM speech detection, using the log step size $d_L(i\tau_0)$ from the lower subband only. $M(i\tau_0)$ is the maximum of the values $\sigma_M d_L(i\tau_0)$; σ_M is a constant weight. Therefore, when $\sigma_M d_L(i\tau_0)$ increases, $M(i\tau_0)$ increases when $\sigma_M dL(i\tau_0)$

decreases, $M(i\tau_0)$ decreases only very slowly according to the leak factor β_M. $M(i\tau_0)$ is restrained from decreasing to less than its lower limit (M_0), so $M(i\tau_0)$ measures the maximum speech energy in the lower subband.

The variable d'_L can be defined to be

$$d'_L(i\tau_0) = d_L(i\tau_0) + 32; \quad (2)$$

the bias of 32 is used to insure that d'_L and M are always positive. The value of M at time $i\tau_0$ is

$$M(i\tau_0) = \max\left\{\beta_M M(i-1)\ \tau_0, \sigma_M d'_L(i\tau_0), M_0\right\} \quad (3)$$

The thresholds are fixed distances below M, so, the threshold λ_{ON}, used to determine when speech changes from OFF to ON, is computed as follows:

$$\lambda_{ON}(i\tau_0) = M(i\tau_0) - C_{ON} \quad (4)$$

the threshold λ_{OFF}, used to determine when speech changes from ON to OFF, is

$$\lambda_{OFF}(i\tau_0) - C_{OFF}; \quad (5)$$

the values of C_{ON} and C_{OFF} are constants, with $C_{OFF} > C_{ON}$.

FIG. 4 shows how the comparison is done. The speech samples are divided into blocks of some convenient length. (In this case 24 samples per block are used.) Once per block, a decision is made concerning whether speech is ON or OFF. If, in the previous block, speech was on, then the ON threshold is used; if speech was off, the OFF threshold is used. The switch in FIG. 4 chooses the correct threshold, which is then compared to the detection statistic. The speech flag is set ON or OFF depending on whether the detection statistic is above or below the threshold. Let τ_{DS} be the time interval associated with one block. (In this case, $\tau_{DS} = 24\tau_0$.) Let S

denote the speech state with two possible values:

$$S = \left\{ \begin{array}{l} 0,\text{-to-indicate-speech-presence,} \\ 1,\text{-to-indicate-silence.} \end{array} \right\} \quad (6)$$

The speech state $S(i\tau_{DS})$ at time $t=i\ \tau_{DS}$ depends on the previous speech state $S[(i-1)\ \tau_{DS}]$ as follows:

when

$S[(i-1)\ \tau_{DS}] = 0.$

$$S(i\tau_{DS}) = \left\{ \begin{array}{l} 0, \text{ if } T(i\tau_{DS}) \geq \lambda_{OFF}\ (\tau_{DS}); \\ 1, \text{if } T(i\tau_{DS}) < \lambda_{OFF}\ (i\tau_{DS}),\ i=1,2,\ldots; \end{array} \right\} \quad (7)$$

when $S[(i-1)\ \tau_{DS}] = 1,$

$$S(i\tau_{DS}) = \left\{ \begin{array}{l} 0, \text{ if } T(i\tau_{DS}) > \lambda_{ON}(i\tau_{DS}); \\ 1, \text{ if } T(i\tau_{DS}) \leq \lambda_{ON}\ (i\tau_{DS}),\ i=1,2,\ldots \end{array} \right\} \quad (8)$$

The system 10 can be effectively implemented by a person of ordinary skill in the art of subband encoding by appropriately adapting two or more digital signal processor microcomputers. Such microcomputers are presently in use and may include a memory unit, an arithmetic unit, a control unit, an input-output unit, and a machine language storage unit in a single VLSI circuit. Their function may alternately be provided by a combination of a number of different VLSI circuits interconnected. One such microcomputer which is suitable for implementing the system 10 is a DSP (Digital Signal Processor) manufactured by AT&T Technologies, Inc., a corporation of New York, U.S.A. and described, for example, in the above-mentioned Bell System Technical Journal volume.

In one example of a system implemented with two DSP's, one DSP is used for the encoding and transmission of

speech, while the other DSP is used for the reception and decoding of speech. External logic is used to interface the PCM (pulse code modulation) bit streams of each DSP to both analog-to-digital and digital-to-analog converters for speech input and output. The DSP microcomputers also perform speech-silence detection on the speech signal, so that the silence intervals can be used to transmit user-supplied data.

The DSP microcomputers determine the speech state every two milliseconds. The transmitting DSP provides the speech-state status for external circuitry and generates a 112-bit frame for transmission. The frame consists of a 3-bit framing pattern, a 1-bit speech flag, and 24 samples of subband encoded speech. This speech is sampled at a 12 Khz rate and encoded with 5-bit accuracy in the low band and 4-bit accuracy in the high band. When the DSP indicates the speech flag is on, external line interface circuitry will send the DSP-generated frame intact. When the speech flag is off, the 24 samples of speech is replaced by 108 bits of user supplied data. After construction, the frame is sent over a 56 Kbps (kilobits per second) digital channel to another terminal for decoding.

In the receiver, a simple framing algorithm is implemented with a combination of DSP firmware and external line interface circuitry. The framing algorithm searches the incoming 56 Kbps signal to find the orientation of the 3-bit framing pattern. After the receiving DSP synchronizes itself with the framing pattern, it reads the speech state flag. If the speech state flag is present, the DSP begins decoding the incoming speech signal for listening, but if the flag is absent, the DSP signals external circuitry to remove the data and send it to a user interface. This pattern is repeated every two milliseconds, as long as a valid framing pattern is detected.

The equations above describe the general concepts involved in determining the quantities needed by the speech

detector. Due to finite bit length and timing considerations in the DSP, some of these equations are preferably slightly modified. For example, the system 10 is based on a 24-sample frame, so every 24 samples a decision is made as to whether speech is present. The speech detection statistic is computed in this framework by the DSP as follows:

$$T(i\tau_{DS})=\beta_{DS}T[(i-1)\tau_{DS}]$$
$$+\sum_{j=1}^{24}\sigma_L d'_L[j\tau_0+(i-1)\tau_{DS}]$$
$$+\sigma_H d'_H[j\tau_0+(i-1)\tau_{DS}] \quad ,j=1,2,... \qquad (9)$$

So $T(i\tau_{DS})$ is updated each sample period by adding $\sigma_L d'_L + \sigma_H d'$ to it, and it is leaked once per block of 24 samples. The value of the maximum level M must also be computed slightly differently to obtain accurate results with the DSP. Let τ_{MAX} be the time interval between two successive points at which M is leaked. Experimentally, it was found that $\tau_{MAX}=8$ seconds works well. The equation for M that may be implemented in the DSP is

$$M(i\tau_0) = \left\{\max \sigma_M d'_L(i\tau_0), M[(i-1)\tau_0]\right\}_{\text{for mod}_{\tau_{MAX}} i\tau_0\neq 0, i=1,2,...},$$
$$M(i\tau_0)= \max\left\{\beta_M \max_{\text{for mod}_{\tau_{MAX}} i\tau_0=0, i=1,2,..}\left\{\sigma_M d'_L(i\tau_0), M[(i-1)\ \tau_0]\right\}, M_0\right\}, \qquad (10)$$

The thresholds only need to be computed once per 24 samples, so that they can be used to detect the presence or absence of speech.

$$\lambda_{ON}(i\tau_{DS})=M(i\tau_{DS})-C_{ON};$$
$$\lambda_{OFF}(i\tau_{DS})=M(i\tau_{DS})-C_{OFF} \qquad (11)$$

The speech state is determined in the same way as described in Section II.2 by equations (6-8).

This invention is not limited to two-band subband coding. The detection statistic $T(i\tau_0)$ and maximum level $M(i\tau_0)$ can include information from a larger number of subbands, using equations similar to equations (1) - (11) above. Silence detection with five-band subband coding is an example of this. Let $d_j(i\tau_0)$ for $j = 1,\ldots, 5$ be the log step size values for each of the five bands, let σ_j, $j = 1,\ldots,5$ be fixed weights, and let β_{DS} be a leak factor slightly less than 1. In analogy with equation (1), a general equation describing the speech detection statistic is

$$T(i\tau_0)=\beta_{DS}T[(i-1)\tau_0]+\sum_{j=1}^{5}\sigma_j d_j(i\tau_0). \qquad (12)$$

Letting μ, $j=1,\ldots,5$ be fixed weights, and β_M a fixed leak factor slightly less than 1, the general equation for the maximum level is

$$M(i\tau_0)=\max\left\{\beta_M M[(i-1)\tau_0],\ \sum_{j=1}^{5}\mu_j d_j(i\tau_0), M_0\right\}. \qquad (13)$$

Some of the weighting factors σ_j or μ could be zero. As in equations (9)-(11), equations (12)-(13) can be slightly altered to conform to a specific hardware implementation, such as an implementation using a DSP microprocessor. It is also necessary to choose specific values of the parameters in equations (12)-(13). For the computation of the detection statistic, $\sigma_1 = \sigma_2$, and $\sigma_3 = \sigma_4 = 2\sigma_1$, giving a greater weight to the higher frequency bands; band 5 is not used, so $\sigma_5 = 0$. For the computation of a maximum level, $\mu_1 = \mu_2$, and $\mu_3 = \mu_4 = \mu_5 = 0$. The maximum level depends on the energy in the low-frequency bands, giving a smooth long-term average.

In theory, the equations (12) and (13) can be

extended to any number of bands. However, as the number of bands increases, the time delay associated with computing the detection statistic and maximum level also increases. Therefore there is a practical limit to the number of bands that can be used in this system.

Claims

1. Signal encoding apparatus
CHARACTERIZED BY
means for encoding a plurality of frequency subband portions of a signal, including means for generating voltage step size values for signal samples of each subband;
means for computing speech statistic values based on the voltage step size values for the one frequency subband and the voltage step size values for another of the frequency subbands; and
means for comparing speech presence energy threshold values and speech silence energy threshold values to the speech statistic values to selectively generate speech presence output signals.

2. The apparatus defined in claim 1 wherein said speech statistic value computing means is
CHARACTERIZED BY
means for multiplying the step size values of each subband by a corresponding speech detection coefficient to generate respective speech detection value products;
means for summing the speech detection value products to generate speech detection value sums, and
means for smoothing the speech detection value sum.

3. The apparatus defined in claim 2
CHARACTERIZED IN THAT
said smoothing means comprises means for summing each speech detection value sum with a delay value to generate a speech detection statistic output value, the delay value being the product of a detection constant and a previous detection statistic output value.

4. The apparatus defined in claim 3
CHARACTERIZED BY
means for computing speech energy threshold values and speech silence threshold values based on the

voltage step size values for one of the subbands.

5. The apparatus defined in claim 4 wherein said speech statistic value computing means is

CHARACTERIZED BY

means for generating a speech presence threshold value and a speech silence value from a maximum energy level value, the maximum energy level value being generated by choosing the maximum of first and second energy levels, the first energy level being the product of a step size value of the low frequency subband and the second energy level being the larger of the previous sample maximum energy level value multiplied by a coefficient and a lower limit.

6. The apparatus defined in claim 5

CHARACTERIZED BY

switch means which connect either the speech threshold value or the speech silence value from the generating means to a one input of a comparator in response to a control signal, the other input of the comparator being connected to receive the speech detection statistic, and

feedback means including a one-sample delay means connected between the output of said comparator and said switch for generating the control signals.

7. A method of detecting the presence of speech content in a signal,

CHARACTERIZED BY

computing a short term speech statistic from the step size value information of at least two of the subbands, and

comparing the speech statistic to a long term speech energy threshold to selectively generate a speech presence indication signal.

8. The method defined in claim 7 further

CHARACTERIZED BY

computing a long term speech energy threshold from the step size information of at least one of the

subbands.

9. The method defined in claim 8 CHARACTERIZED BY

giving greater weight to the step size values for a higher frequency subband than to those of a lower frequency subband when computing the short term speech statistic.

0167364

FIG. 1

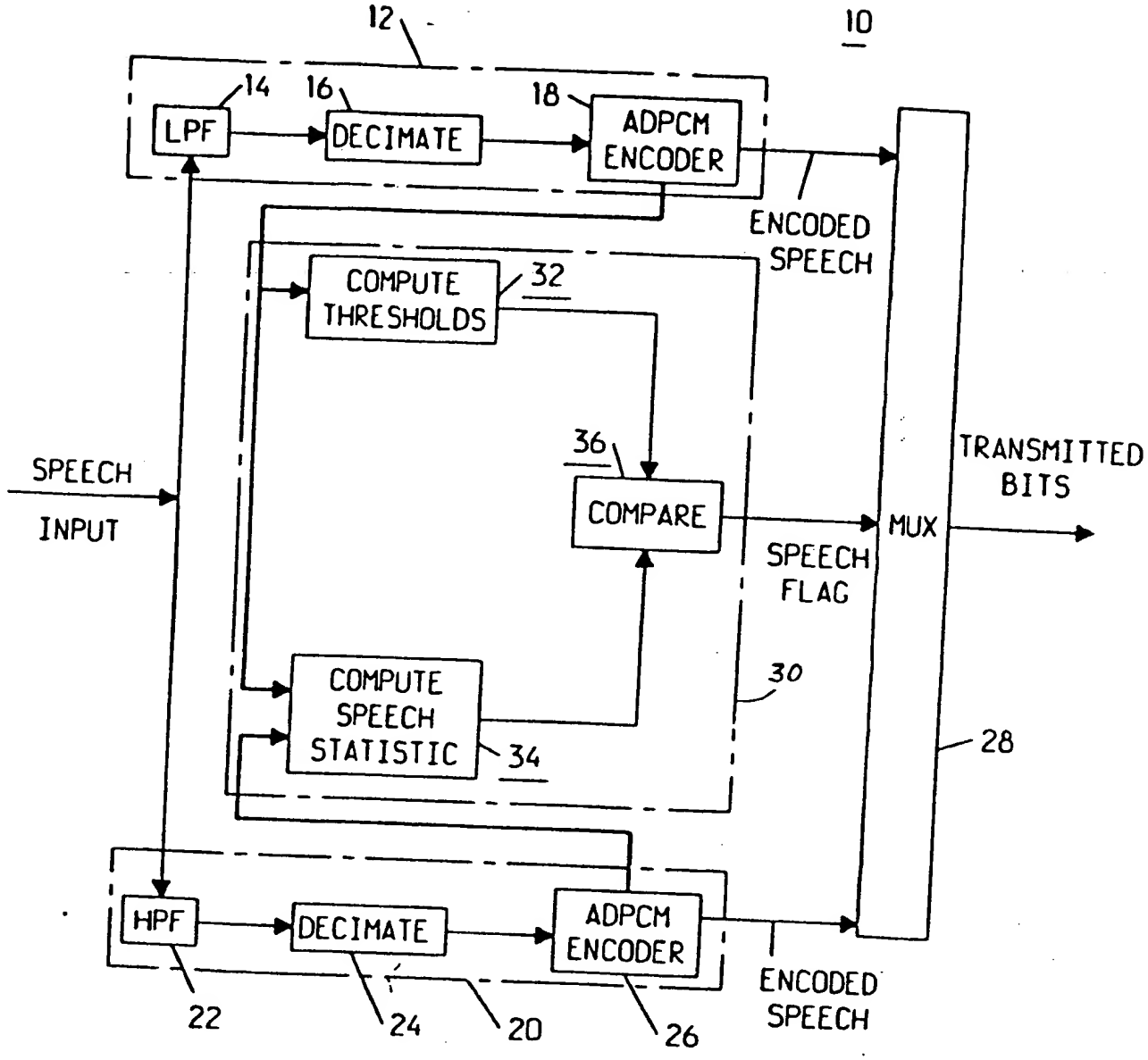


FIG. 2

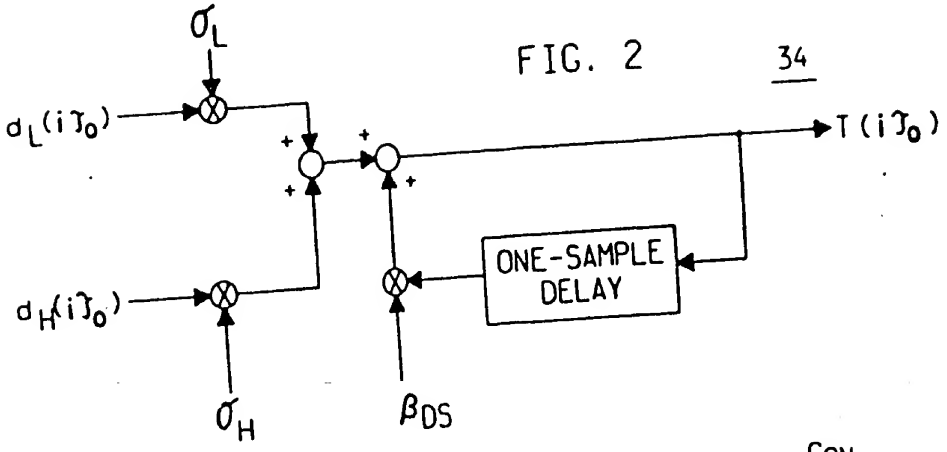


FIG. 3

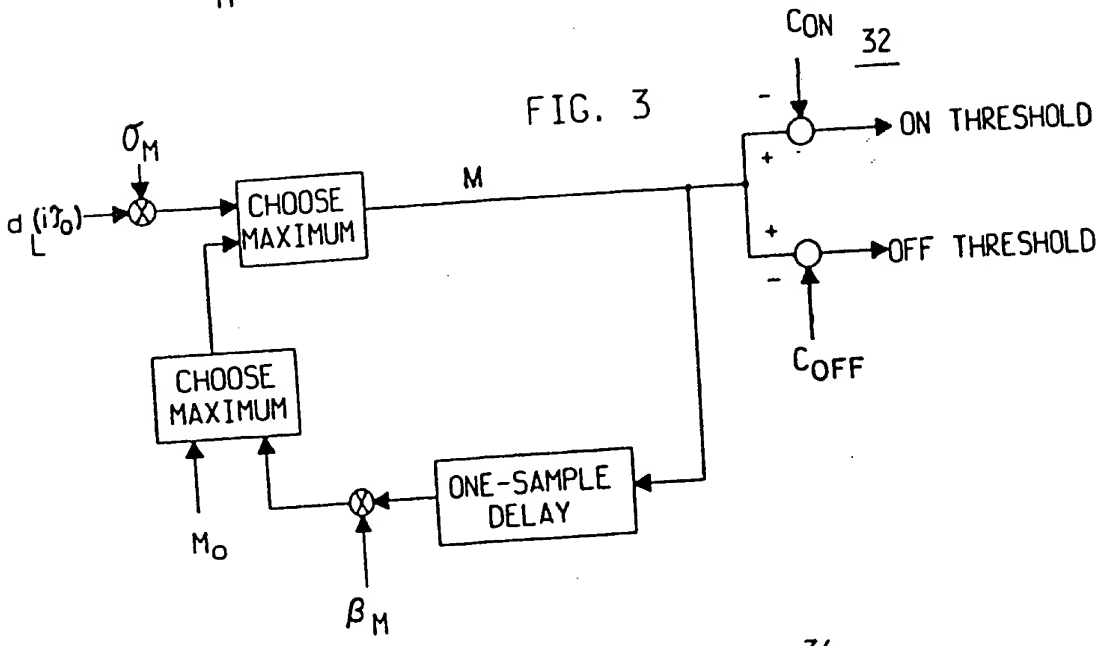
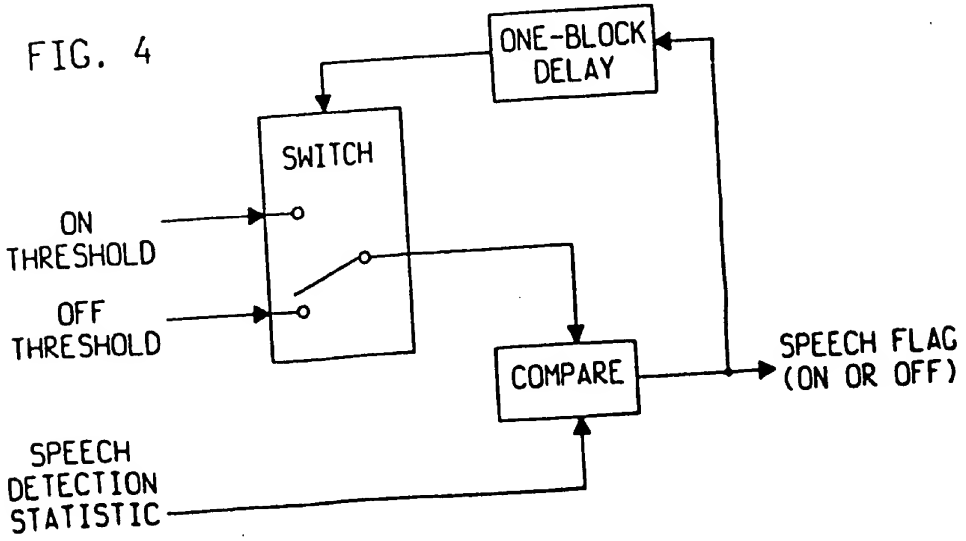


FIG. 4



European Patent Office

EUROPEAN SEARCH REPORT

Application number 0167364

EP 85 30 4627

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. Cl.4)
A	IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, vol. ASSP-28, no. 5, October 1980, pages 550-561, IEEE, New York, US; B.V. COX et al.: "Nonparametric rank-order statistics applied to robust-voiced-unvoiced-silence classification" * Paragraph V: "System Description" *	1	G 10 L 3/00
A	TELECOMMUNICATIONS AND RADIO ENGINEERING, vol. 4, April 1965, pages 70-72, Washington, US; V.N. TETEREV: "A combinatorial method of detecting speech signals in a background of smooth noise" * Figures 2,3 *	1	
A	EP-A-0 110 467 (PHILIPS KOMMUNIKATIONS INDUSTRIE AG et al.) * Claims 3,4 *	3	TECHNICAL FIELDS SEARCHED (Int. Cl.4) G 10 L 3/00
A	DE-A-3 235 279 (NISSAN MOTOR) * Claim 14 *	3	
A	FR-A-2 451 680 (J. SOUMAGNE) * Claim 1 *	5	

-/-

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	20-08-1985	ARMSPACH J.F.A.M.

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO Form 1503 03.82

European Patent Office

EUROPEAN SEARCH REPORT

0167364
Application number

EP 85 30 4627

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. Cl.4)
A	IEEE TRANSACTIONS ON COMMUNICATIONS, vol. COM-24, no. 5, May 1976, pages 563-567, New York, US; R.W. SCHAFER et al.: "Detecting the presence of speech using ADPCM coding" * Abstract * -----	1	
			TECHNICAL FIELDS SEARCHED (Int. Cl.4)

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	20-08-1985	ARMSPACH J.F.A.M.

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO Form 1503 03.82